Review

Not peer-reviewed version

Artificial Intelligence in Systematic Reviews: Overcoming Reproducibility, Bias and Validation Challenges

Mario Treviño * and Oscar Arias-Carrión *

Posted Date: 24 June 2025

doi: 10.20944/preprints202506.1895.v1

Keywords: Artificial Intelligence; Machine Learning; Systematic Reviews; Evidence-Based Medicine; Reproducibility; Algorithmic Bias; Transparency; Information Retrieval

Review

Artificial Intelligence in Systematic Reviews: Overcoming Reproducibility, Bias and Validation Challenges

Mario Treviño [1,*,†] and Oscar Arias-Carrión [2,3,*,†]

[1] Laboratorio de Plasticidad Cortical y Aprendizaje Perceptual, Instituto de Neurociencias, Universidad de Guadalajara, México

[2] División de Neurociencias | Clínica. Instituto Nacional de Rehabilitación Luis Guillermo Ibarra Ibarra. Mexico City 14389, Mexico

[3] Tecnologico de Monterrey, Escuela de Medicina y Ciencias de la Salud. Mexico City 14380, Mexico

* Correspondence: mariomtv@hotmail.com (M.C.); ariasemc2@gmail.com (O.A.-C.)

† These authors contributed equally to this work.

Abstract

Artificial Intelligence (AI) is rapidly changing how systematic reviews are conducted by accelerating the processes of literature retrieval and screening. While these advancements enhance researchers' productivity, the complete scope of AI's transformative potential is still emerging. Moreover, issues related to reproducibility, bias, and transparency pose significant barriers to fully integrating AI into evidence synthesis. Large language models and machine learning classifiers show high sensitivity but suffer from low specificity, generating excessive false positives that increase the screening burden rather than reducing it. AI-generated Boolean search strategies often lack stability, frequently delivering inconsistent results for the same prompts, which undermines the core principle of reproducibility. Furthermore, AI models can sometimes "hallucinate," a term used to describe instances where the AI generates false or misleading information. They may also misapply Medical Subject Headings (MeSH) and introduce selection bias, ultimately distorting the outcomes of systematic reviews. This review examines the role of AI in systematic searching and presents a structured validation framework to address these limitations. Establishing standardized benchmarks for reproducibility, managing sensitivity and specificity trade-offs, and developing clear explanatory mechanisms are crucial to ensure that AI is a complementary tool in evidence synthesis, rather than a disruptive force. Retrieval-augmented AI search frameworks can improve precision but require transparent decision-making processes to enhance trust and accountability. Hybrid AI-human workflows, where AI accelerates screening but human experts validate outputs, offer a pragmatic solution to balance efficiency with methodological rigor. This review presents a comprehensive roadmap emphasizing the importance of interpretability, transparency reports, and ethical oversight to facilitate the responsible integration of AI into systematic reviews. Achieving reproducibility and reducing bias is critical for transforming AI from an experimental tool into a more reliable asset for evidence synthesis.

Keywords: artificial intelligence; machine learning; systematic reviews; evidence-based medicine; reproducibility; algorithmic bias; transparency; information retrieval

1. Introduction

Artificial intelligence has the potential to revolutionize systematic reviews and evidence-based medicine by enhancing the processes of literature retrieval and screening, making them faster and more efficient. While its full range of capabilities is still evolving, unresolved reproducibility, transparency, and reliability challenges impede AI's integration into evidence synthesis [1-6]. Large language models and machine learning algorithms excel at capturing relevant studies with high sensitivity, meaning they have a strong ability to identify true positives (i.e., studies that are correctly identified by AI models as appropriate to the research question). However, their low specificity can lead to false positives (i.e., studies that are incorrectly identified as relevant when they are not), creating an increased workload for manual screening rather than alleviating it (**Figure 1**). AI-generated Boolean search strategies can exhibit instability and non-reproducibility, producing inconsistent outputs even under identical search conditions. Research has shown that AI models often "hallucinate" search terms, producing information that seems plausible but is incorrect or nonsensical. For example, this occurs when AI suggests search terms that don't align with the intended research criteria, potentially leading to irrelevant or misleading results. Additionally, these models may misapply Medical Subject Headings (MeSH) and generate incomplete or misleading queries, which can introduce selection bias and potentially distort evidence synthesis. The role of retrieval-augmented AI search frameworks in mitigating these issues remains promising but unproven, as opacity in AI decision-making complicates efforts to ensure trust, reproducibility, and accountability in systematic searching.

Recent evaluations of large language models in systematic reviews have highlighted critical weaknesses in the trade-off between sensitivity and specificity [5,7,8]. For instance, studies comparing human and AI-assisted screening report that AI models achieve up to 93% sensitivity but at the cost of specificity reductions as low as 25%, leading to an overwhelming number of false-positive selections [9] (**Table 1**). Moreover, the variability in AI-generated search outputs challenges the fundamental principle of reproducibility in systematic reviews. Indeed, identical queries can yield divergent results due to subtle changes in prompts [10]. This instability highlights the need for standardized AI-assisted methodologies and validation benchmarks to ensure the reliability and rigor of systematic reviews [11,12].

Despite these limitations, AI has shown efficiency gains in tasks such as title and abstract screening. Tools like Rayyan and EPPI-Reviewer reduce manual workload by prioritizing relevant studies, although human oversight remains essential [13]. Integrating hybrid AI-human workflows, where AI assists in preliminary screening while experts validate selections, has been proposed as a pragmatic solution to enhance efficiency without compromising methodological rigor [3].

This review proposes a hybrid AI-human model for systematic searching, integrating rigorous validation frameworks, standardized AI-assisted methodologies, and benchmarked reproducibility metrics to enhance reliability. It emphasizes the need for dedicated AI reporting standards, the refinement of sensitivity-specificity trade-offs, and the creation of interpretable models to ensure AI enhances, rather than hinders, evidence synthesis (**Table 1**). This review offers a structured roadmap for the responsible integration of AI into systematic reviews, moving beyond mere automation to prioritize a balanced approach that combines efficiency with methodological rigor, transparency, and long-term sustainability.

Table 1. Artificial intelligence in systematic reviews: key applications, models, and insights.

Study	Aim and Focus	AI Model(s)	Main Task(s)	Key Findings	Critical Insight
O'Connor et al. (2019) [15]	Trust and set up barriers in SR automation	Conceptual model	Workflow, adoption	Trust hinges on compatibility with current practice	Human workflows resist full automation unless seamlessly integrated

Staudinger et al. (2024) [16]	Query reproducibility across LLMs	GPT-3.5, Alpaca, Flan-T5	Query generation	Manual refinement” auto-generation	Human-LLM hybrid outperforms automation alone
Badami et al. (2023) [2]	Boolean query generation & refinement	Transformer LM	Query → refinement → retrieval	Adaptive models improve recall with manual input	Human-in-the-loop remains essential for precision
Tercero-Hidalgo et al. (2022) [17]	AI use in COVID-19 literature reviews	ML classifiers	Search, screening	Improved speed and coverage with AI filtering	Feasible in time-sensitive contexts, with oversight
Feng et al. (2022) [6]	AI meta-analysis on literature screening	Various ML/AI	Screening	AUC = 0.86; high sensitivity, medium precision	Suitable for sensitivity; specificity requires tuning
Guo et al. (2024) [7]	GPT-4 vs humans in paper screening	GPT-4 API	Title and abstract screening	κ = 0.96 (vs humans); sensitivity = 0.76–0.91	GPT-4 can augment humans in clinical review tasks
Wilkins (2023) [18]	GPT-4 for scoping review screening	GPT-4	Abstract screening	Agreement = 84% vs humans; specificity > sensitivity	GPT is a feasible co-pilot, not a replacement
Khraisha et al. (2024) [19]	GPT-4 in real-world SR	GPT-4	Screening	High recall, false positives ↑ with prompt drift	Promising, but context and tuning matter
Li et al. (2024) [20]	Compare LLMs (PaLM, Claude, etc.) in SR	ChatGPT, PaLM, Claude, LLaMA	Abstract screening	LLMs ≈ humans in some metrics; performance varies	Fine-tuning needed per domain/task
Abogunrin et al. (2025) [21]	Cost-benefit of AI in SR workflows	ML, NLP tools	Efficiency, cost, recall	WSS@95 up to 60%; time ↓, human hours saved	AI can significantly cut costs with minimal loss
Matsui et al. (2024) [22]	3-layer GPT workflow for screening	GPT-3.5, GPT-4	Screening	Sens = 0.962; Spec = 0.996 (GPT-4)	Prompt structuring is key to high-precision
Fenske & Otts (2024) [23]	Teaching AI use via Elicit in nursing	Elicit	Search training	Students ↑ in critical thinking and aided inquiry	Good pedagogy; not validated for SR use
Waffenschmidt et al. (2023) [24]	Efficiency via prioritization tools	EPPI, Rayyan	Screening prioritization	EPPI + single screening missed a few studies	Time-saving is feasible with safeguards
Lieberum et al. (2025) [25]	LLMs across 10/13 SR stages	GPT-3.5, ChatGPT	End-to-end SR mapping	54% “promising”, 22% “non-promising”	Broad use, little standardization or validation
Adam et al. (2024) [26]	Boolean query automation with LLMs	Mistral-Instruct-7B	Query generation	Sens = 85%; NNT = 1206	Draft tool only; librarians are still essential

Cierco Jimenez et al. (2022) [4]	Map 63 ML tools for SR stages	SVM, classifiers	Tool mapping	Tools exist; they lack validation and usability	Workflow integration remains a barrier
Thomas et al. (2021) [27]	ML classifiers in EPPI-Reviewer	Cochrane, EPPI tools	Screening prioritization	Recall = 99%; classification optimized	Domain-trained models boost precision
Siemens et al. (2025) [28]	Risks and benefits of AI in SR	LLMs, ML tools	SR stages	Data extraction > search for reliability	AI ≠ replacement: High-stakes reviews need humans
Markolf et al. (2021) [29]	AI in infrastructure complexity	General AI types	Theoretical framework	Align AI levels to system complexity	No SR-specific tools; broader governance warning
Park & Thomas (2018) [30]	ML functionalities in the EPPI platform	EPPI classifiers	Classification, prioritization	≥99% recall; real-time feedback	Widely usable, transparent and validated
Oami et al. (2024) [9]	GPT-4 Turbo in sepsis review	GPT-4 Turbo	Screening	Sens = 0.91, Spec = 0.98; time ↓	Reliable in clinical domains with tuning
Tran et al. (2024) [10]	GPT-3.5 prompt tuning for SRs	GPT-3.5 Turbo	Screening	Sens up to 99.8%, Spec variable	Trade-off: recall ↑, precision ↓
Parisi & Sutton (2024) [14]	ChatGPT for search query dev	ChatGPT	Search formulation	Mostly opinion-based, validation lacking	Not ready for SR use; oversight mandatory
Khalil et al. (2022) [13]	Review automation tools	Rayyan, DistillerSR	Screening, extraction	Screening tools mature; others lag	Tool usability is uneven
Delgado-Chaves et al. (2025) [5]	LLM comparison across 3 SRs	18 LLMs	Screening	Workload ↓ 33–93%; outcome-sensitive	Task framing is critical to performance
Blaizot et al. (2022) [3]	AI in health sciences SRs	15 AI models	Screening, extraction	73% used in screening; all human-verified	Full automation is not yet viable
Page et al. (2021) [31]	PRISMA 2020 explanation	None	Reporting	27-item checklist updated	The benchmark for AI-assisted SR transparency

2. AI-Assisted Systematic Searching

2.1. The promise and limitations of AI-generated Boolean queries

AI is increasingly positioned as a transformative tool in systematic searching, with large language models such as ChatGPT, Mistral, and Claude capable of generating Boolean search strategies within seconds [14]. By interpreting natural language inputs and converting them into structured search commands, these models offer the potential to improve the traditionally labour-intensive query formulation process. Their ability to process vast amounts of text and rapidly generate search strings suggests a radical paradigm shift in evidence synthesis, reducing the need for manual construction of complex Boolean queries.

doi:10.20944/preprints202506.1895.v1

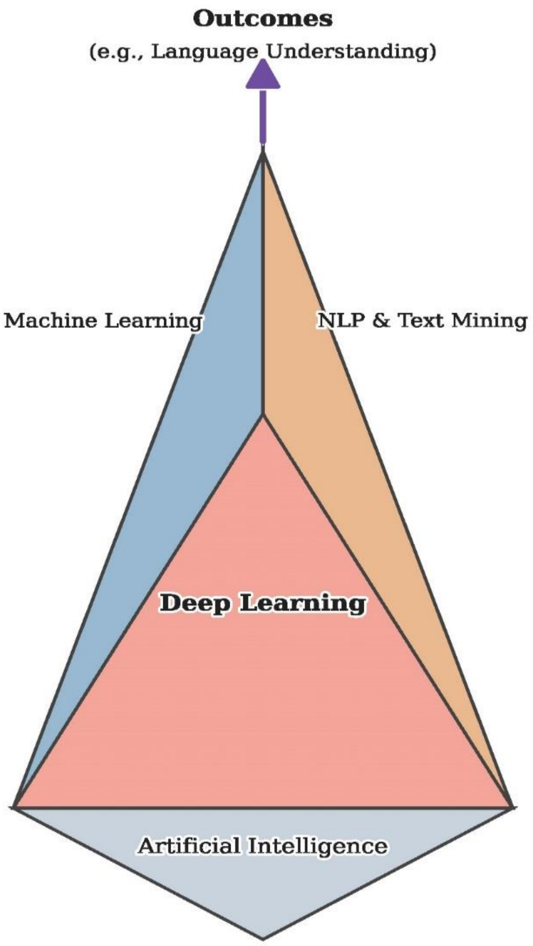

Figure 1. Convergence of artificial intelligence, machine learning, and natural language processing in systematic searching. Artificial intelligence (AI) encompasses a spectrum of computational tools that can automate literature retrieval. Machine learning (ML), a branch of artificial intelligence (AI), employs supervised and unsupervised learning to enable predictive filtering and classification. Deep learning (DL), nested within ML, powers large language models (LLMs) capable of semantic reasoning (the ability of a model to process the meaning and context of words and phrases) and Boolean query optimization. Natural language processing (NLP) supports these processes by enabling entity recognition, term normalization, and the extraction of structured information from text. At the intersection of DL and NLP, LLMs represent a transformative development in automated systematic searching, offering novel pathways for scalable and context-sensitive evidence synthesis.

Despite this promise, AI-generated Boolean queries lack a critical feature of systematic search: reproducibility. Identical prompts can yield variable outputs, even when controlling for database specifications, making it difficult to verify or replicate search results. This inconsistency undermines a foundational principle of systematic searching, where transparent and repeatable methods are essential for ensuring validity. Furthermore, AI models frequently introduce errors by fabricating search terms—“hallucinating” nonexistent keywords or misattributing Medical Subject Headings (MeSH)—which can lead to the retrieval of irrelevant literature or the omission of key studies. These inaccuracies introduce a methodological risk, as flawed search queries bias results, compromise sensitivity, and distort the synthesis of available evidence [14] (**Figure 2**). Bias in large language model outputs also stems from their training on opaque datasets that may underrepresent certain research areas or overemphasize specific viewpoints. The reliance on non-transparent training data limits the reproducibility of AI-assisted searches and raises concerns about the inclusion of incomplete or misleading information in systematic reviews. Furthermore, frequent model updates mean that AI-generated search queries may not yield the same results over time, hindering, if not entirely preventing, validation and verification efforts [10].

Addressing these challenges has led to the development of adaptive search query refinement techniques, where iterative human feedback refines AI-generated searches, enhancing precision and stability. While such methods can potentially improve search accuracy, they remain heavily dependent on expert validation and manual correction. Rather than functioning as autonomous tools, AI-assisted Boolean search strategies require continuous human oversight to ensure methodological rigor. This reinforces the need for a hybrid model in which AI augments, rather than replaces, expert-driven systematic searching [3] (**Figure 3**).

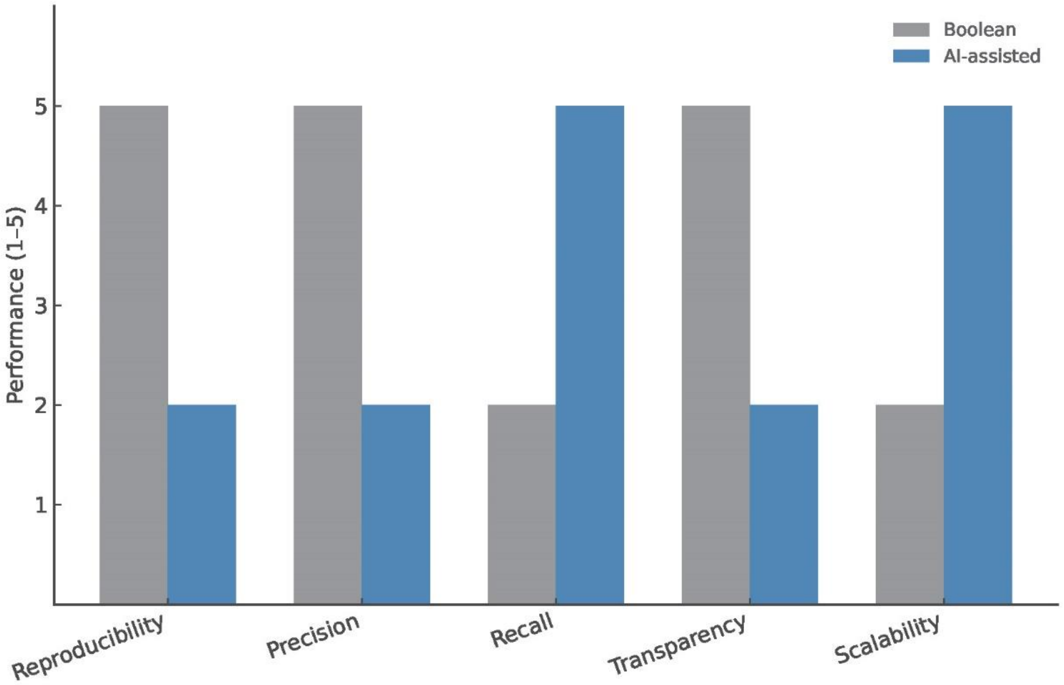

Figure 2. Comparative strengths and limitations of Boolean versus contemporary AI-assisted searching. Boolean-based queries provide reproducible, interpretable, and particular outputs but require manual updating and are constrained by rigid syntax, limiting adaptability and recall. In contrast, AI-assisted searches are scalable and can retrieve broader evidence bases but suffer from reduced specificity, lower reproducibility, and opaque decision-making. Reproducibility, precision, recall, transparency, and scalability metrics reveal the complementary strengths of Boolean and AI-assisted search methods. Combining AI-driven query generation and expansion with human validation and filtering can enhance the rigor and efficiency of systematic reviews.

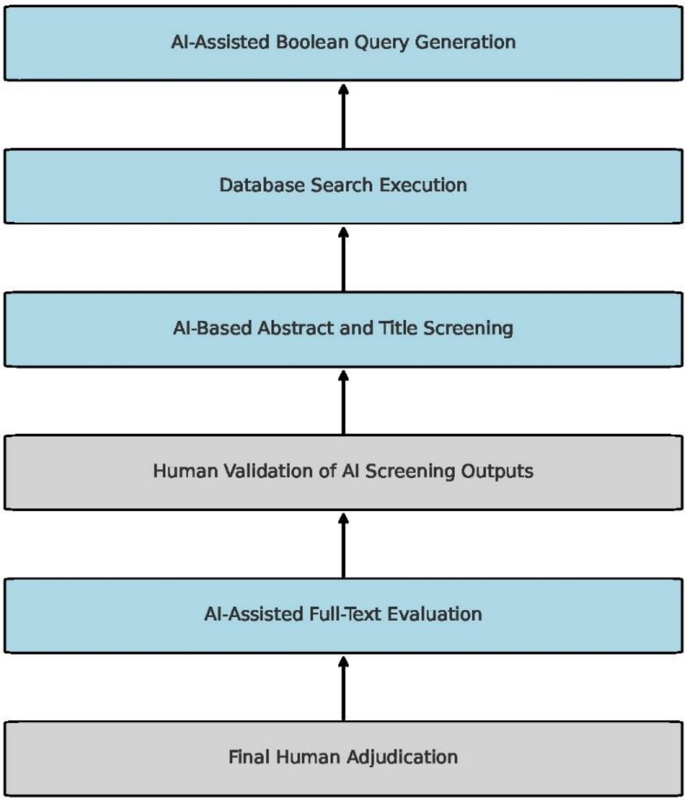
AI-Assisted Boolean Query Generation
Database Search Execution
AI-Based Abstract and Title Screening
Human Validation of AI Screening Outputs
AI-Assisted Full-Text Evaluation
Final Human Adjudication

Figure 3. AI-driven systematic screening workflow with embedded human oversight. This flowchart illustrates a staged hybrid pipeline for AI-assisted systematic screening. The initial steps involve a large language model, which generates Boolean queries tailored to database specifications, followed by automated query execution across repositories. AI tools then rank abstracts and titles, prioritizing studies based on relevance. Expert review ensures methodological fidelity at key checkpoints by validating AI outputs, filtering false positives, and confirming study eligibility. Natural language processing tools aid in full-text evaluation and bias assessment, but final adjudication—encompassing clinical relevance, synthesis, and interpretation—remains the exclusive domain of humans. This layered workflow aims to maintain scientific rigor while reducing the screening burden.

2.2. Precision vs. recall: AI's fundamental trade-off

AI-based search tools promise to enhance efficiency in evidence synthesis; however, their performance often highlights an inherent trade-off between precision (avoiding irrelevant results) and recall (capturing all relevant results; **Table 3**). By prioritizing recall, these models aim to capture the broadest possible set of appropriate studies, thereby minimizing the risk of omission. However, this approach comes at a cost—diminished precision—resulting in an influx of false positives that can overwhelm researchers with an excessive screening burden [9]. Unlike traditional Boolean searches, which rely on meticulously crafted logic to refine specificity, AI-generated queries operate with a more expansive retrieval strategy, frequently sacrificing control over search refinement in favor of inclusivity.

Efforts to mitigate AI's low specificity have led to adopting retrieval-augmented generation (RAG) models, which integrate real-time literature retrieval into AI-generated responses. These models enhance contextual relevance by pulling text from existing databases before formulating search queries, allowing AI to refine results based on dynamically retrieved information rather than relying solely on static training data. While this approach improves precision, its effectiveness remains contingent on the quality of the initial search prompt. Even with real-time augmentation, RAG models inherit the limitations of their inputs, reinforcing the need for expert oversight in query formulation [5].

Despite these advancements, AI's reliance on machine-learned relevance scores, rather than systematic logic, in ranking search results complicates its integration into systematic searching. While studies suggest that AI-powered search tools achieve high recall, their specificity varies widely (25–80%), resulting in substantial inefficiencies in screening. This fundamental trade-off between comprehensive recall and manageable specificity remains a central challenge in AI integration into systematic review methodologies.

3. Literature Retrieval: The Myth of Full AI Automation

AI-driven literature retrieval tools, such as Elicit, ResearchRabbit, and Consensus, claim to accelerate the initial phases of systematic searching by replacing traditional Boolean logic with natural language queries, enhancing efficiency and accessibility [3]. By offering an intuitive and user-friendly interface, these platforms promise to automate search processes, reducing the need for manual query formulation. However, their reliance on open-access sources and pre-indexed content limits their ability to comprehensively retrieve relevant studies. Unlike traditional search engines that provide real-time access to proprietary databases such as PubMed, Cochrane, and Embase, AI-assisted retrieval tools are often constrained by paywalls and subscription-based repositories, frequently missing key studies critical for a systematic review [13].

Beyond incomplete coverage, the opaque decision-making processes of AI-based retrieval systems further complicate their role in evidence synthesis. Unlike Boolean searches, where users can trace and refine search logic with complete transparency, AI-generated searches function as black boxes, offering little insight into why specific papers are included or omitted [14]. This lack of explainability raises concerns about selection bias, as AI models may favor highly cited papers over emerging research, further skewing systematic review findings [5]. Without clear inclusion criteria or transparency in search logic, AI-based retrieval tools remain supplementary rather than standalone solutions, underscoring the necessity of human oversight in systematic searching (**Table 2**).

Moreover, the environmental impact of AI-driven research must also be considered. LLM-powered retrieval systems require substantial computational power, consuming four to five times more energy than traditional database queries [9]. This increased energy consumption contributes to sustainability concerns, further challenging the large-scale deployment of AI in systematic reviews.

While AI presents a promising avenue for accelerating systematic searching, its integration into evidence synthesis remains fraught with reproducibility, specificity, selection bias, and transparency challenges. The current limitations of AI-powered searching suggest that it should be viewed as a complementary tool rather than a replacement for expert-driven methodologies. The hybrid AI-human approach, where AI accelerates initial searches while human researchers refine and validate the outputs, represents the most viable path forward [10].

Table 2. Key AI tools for systematic searching and screening.

AI Tool	Primary Function	Strengths	Limitations
ChatGPT/Mistral	AI-driven Boolean query generation	Adaptive and flexible	Lack of reproducibility, hallucinated MeSH terms
Rayyan	Abstract and title screening	High sensitivity, semi-automated workflow	Low specificity, false positives
EPPI-Reviewer ML	Machine learning for living systematic reviews	Continuous updates, automation	Requires human validation

Elicit/Consensus	Literature retrieval and ranking	Fast, user-friendly interface	Limited access to proprietary databases
GPT-4 for Risk of Bias	AI-assisted bias assessment	Identifies methodological flaws	Less reliable than trained human reviewers

4. AI for Systematic Screening: Does It Actually Work?

4.1. AI-powered abstract and title screening

AI is increasingly being integrated into systematic review workflows to expedite title and abstract screening, aiming to reduce the manual burden of study selection. AI-driven platforms like Rayyan, AbstrackR, and SWIFT-Active Screener employ machine learning classifiers to prioritize relevant studies while filtering out those unlikely to meet the inclusion criteria [3]. These tools have high sensitivity, capturing up to 96% of pertinent literature, yet their limited specificity leads to an overwhelming number of false positives [9]. As a result, researchers manually review hundreds of irrelevant papers, which diminishes the efficiency gains AI intends to provide.

The limitations of AI-based screening become more evident in real-world applications, where models frequently misclassify 10–20% of studies—a rate higher than acceptable for systematic reviews [10]. AI struggles with subtle eligibility criteria, unlike human reviewers, particularly when study selection depends on methodological details, population characteristics, or intervention specificity rather than simple keyword matching. Even sophisticated large language models lack the contextual reasoning to differentiate between subtle inclusion and exclusion factors, leading to inconsistencies that demand human oversight [14].

Another critical challenge lies in AI's potentially biased output resulting from non-transparent training data and variable model performance. Studies suggest that AI screening tools are highly sensitive but suffer from poor reproducibility, as they generate different screening results when applied to the same dataset at other times [5]. The lack of stability in AI-generated screening recommendations raises concerns about consistency, making human intervention necessary to ensure reliable study selection.

Efforts to enhance AI-assisted screening have focused on refining preprocessing techniques, where models are trained on clearly defined inclusion-exclusion parameters before screening begins. These optimizations have shown incremental improvements in specificity, thereby reducing false positives and enhancing efficiency [13]. However, despite these refinements, AI remains a supplementary tool rather than a standalone solution. Edge cases, ambiguous abstracts, and AI-generated errors still require expert validation, reinforcing the need for a hybrid approach in which AI augments rather than replaces human judgment in systematic review methodologies.

Table 3. AI vs. traditional systematic searching – performance comparison.

Metric	AI-Based Searching	Boolean-Based Searching
Reproducibility	Low (variable outputs)	High (consistent results)
Precision	Moderate (high false positives)	High (strict inclusion criteria)
Recall	High (captures broad literature)	Moderate (risk of missing relevant studies)
Transparency	Low (black-box models)	High (fully defined query logic)
Bias Risk	High (citation bias, hallucinated search terms)	Low (controlled search strategy)
Scalability	High (adapts to new data)	Low (manual refinement required)

4.2. *AI in full-text screening and risk of bias assessment*

AI has proven effective in processing large volumes of abstracts, yet full-text screening presents a more complex challenge. Unlike title and abstract screening, where inclusion criteria are relatively straightforward, full-text evaluation demands nuanced judgment of methodological rigor, statistical validity, and domain-specific relevance [11,12]. AI models struggle with these more complex analytical tasks, often failing to detect study quality issues, assess the risk of bias, or identify methodological limitations with the same precision as trained experts. As a result, their role in final inclusion decisions remains limited, reinforcing the need for human oversight in systematic reviews [3].

Attempts to automate risk-of-bias assessments using large language models such as ChatGPT and GPT-4 have yielded mixed results. While these models can recognize overt methodological flaws, their ability to detect subtle biases in study design and statistical reporting remains inferior to that of experienced reviewers [9]. AI lacks the contextual understanding required to evaluate clinical relevance, interpret methodological weaknesses, and identify domain-specific biases—factors essential for ensuring the integrity of systematic reviews.

Additionally, AI-assisted risk-of-bias assessments show only moderate agreement (41–71%) with expert reviewers, suggesting that large language models cannot yet replace human judgment in this critical process [10]. While some studies report higher agreement levels (up to 80% for specific tasks), this remains insufficient for high-stakes decision-making in evidence synthesis. Given the inherent variability in large language model performance, researchers are advised to use AI for preliminary bias screening, with final assessments remaining firmly under expert control [14].

Efforts to refine AI-assisted bias assessment continue through adaptive learning and expert-informed training, yet no current system meets the methodological standards necessary for full automation. AI may assist by highlighting potential concerns or flagging inconsistencies, but full-text screening and bias assessment still require human judgment [5]. Furthermore, the black-box nature of AI models prevents transparency in how bias assessments are conducted, further reinforcing the need for expert oversight.

Beyond technical concerns, AI-assisted screening raises broader issues, including data privacy risks and sustainability challenges. Large language model-powered screening tools consume significantly more computational resources than traditional screening methods, raising concerns about the environmental impact of large-scale AI deployment [13]. These factors underscore the importance of responsible AI integration, ensuring that its adoption does not compromise the transparency, reproducibility, and sustainability of systematic reviews.

While AI-driven systematic screening offers significant efficiency gains, its specificity, bias assessment, and contextual reasoning limitations preclude full automation. The most viable approach remains a hybrid AI-human model, where AI accelerates preliminary screening while expert reviewers maintain final oversight [11,12]. Refining AI-assisted methodologies, increasing transparency, and addressing sustainability challenges will be key to ensuring the responsible integration of AI in systematic review processes.

5. Unresolved Challenges: Reproducibility, Bias, and Sustainability

5.1. *Reproducibility crisis in AI-driven searching*

AI-driven systematic searching faces a fundamental challenge: its lack of reproducibility, a cornerstone of rigorous evidence synthesis. Despite the complexity of traditional Boolean searches, they offer full transparency and replicability, enabling researchers to track, refine, and validate queries precisely. In contrast, large language models generate inherently unstable search strategies, producing different Boolean queries in response to identical prompts [10]. This variability arises from model updates, token randomness, and API versioning, making it impossible to ensure that identical searches yield the same results over time [14].

The instability of AI-generated queries compromises the integrity of systematic reviews, where reproducibility is essential for maintaining methodological rigor (**Figure 4**). Without the ability to replicate AI-generated searches, systematic reviews risk losing their reliability, as different researchers—or even the same researchers at other times—may retrieve varying sets of studies using identical search terms [5].

Beyond the unpredictability of AI-driven queries, the absence of standardized evaluation frameworks further erodes reproducibility. Traditional systematic searches rely on established precision-recall metrics and peer validation to ensure accuracy and completeness. In contrast, no universally accepted benchmark exists for assessing the reliability of AI-assisted search strategies. The lack of clear validation metrics makes it difficult to determine whether AI-generated searches are systematic, exhaustive, or valuable for evidence synthesis [11,12].

While efforts to refine AI-assisted search methodologies continue, including adaptive learning and real-time search augmentation, AI remains a tool requiring human oversight rather than a fully autonomous replacement for expert-driven searching. Without greater transparency and standardized evaluation, AI-assisted systematic searches cannot be relied upon as a standalone methodology [9].

5.2. Bias, black-box AI, and the trust problem

Bias remains a persistent challenge in AI-assisted systematic reviews arising from data limitations and the opaque nature of AI decision-making. AI search and screening tools often exhibit citation bias, disproportionately favoring highly cited studies, established journals, and mainstream research while overlooking emerging literature and underrepresented findings [3]. This bias risks skewing systematic reviews toward dominant narratives, potentially excluding novel, high-impact discoveries from smaller research groups or less visible sources.

Unlike traditional search strategies, which allow researchers to refine inclusion criteria meticulously, AI-driven tools operate probabilistically, making selection decisions that users cannot fully interpret or adjust. Studies suggest that AI-generated inclusion and exclusion decisions are often irreproducible, with large language models retrieving different papers based on minor variations in search prompts or model parameters [13].

The black-box nature of AI further compounds these challenges, limiting transparency in the search and screening process. Boolean queries provide a clear, trackable logic that researchers can audit and refine, whereas AI-generated searches rely on non-explicit algorithms, obscuring the reasoning behind study selection [5]. This lack of explainability hinders trust in AI-assisted systematic searches, especially when study inclusion or exclusion lacks clear justification.

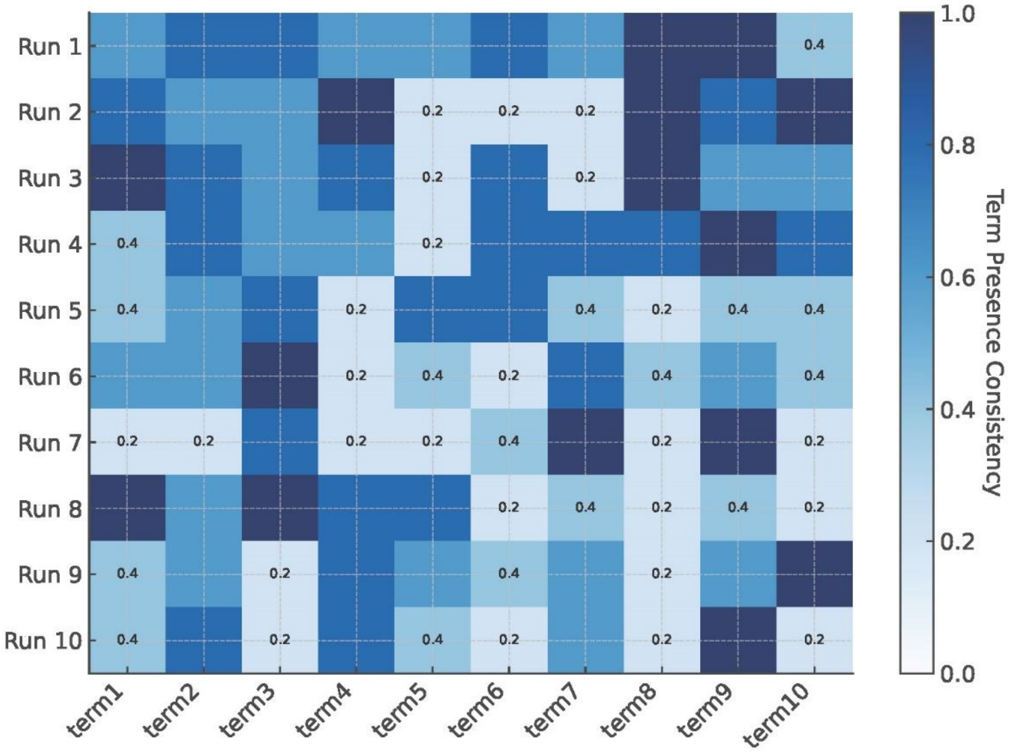

Figure 4. Reproducibility crisis in AI-generated Boolean querying. Heatmap representation of AI-generated Boolean queries across ten identical prompt iterations reveals significant variation in term inclusion. Darker cells indicate stable query terms retrieved consistently, while lighter cells denote unstable terms that fluctuate across runs. These inconsistencies highlight a critical challenge in AI-driven evidence retrieval: reproducibility. Unlike rule-based Boolean strategies, AI-generated queries are influenced by stochastic processes within large language models, undermining transparency and reliability. Standardized validation frameworks are essential for monitoring and mitigating variability in AI-based searching, ensuring systematic reviews' robustness.

Even attempts to enhance AI transparency—such as Retrieval-Augmented Generation (RAG) models—fail to resolve the trust problem, as AI-generated retrievals often rely on highly ranked sources rather than unbiased, comprehensive literature screening [10]. Therefore, without mechanisms to make AI decision-making more interpretable, skepticism surrounding its reliability and fairness persists, reinforcing the need for human oversight to validate AI-driven outputs. Until AI models demonstrate greater transparency, methodological consistency, and explainability, they remain augmentative rather than autonomous tools in evidence synthesis [14].

5.3. The hidden cost: environmental and computational trade-offs

While AI promises to accelerate systematic reviews, its environmental and computational costs remain largely overlooked. Training and deploying large language models for systematic searching requires more energy than traditional database queries, with estimates suggesting a four- to fivefold increase in energy consumption [9]. Unlike Boolean searches, which are optimized for database efficiency and computational simplicity, AI-driven searches rely on millions of parameters and real-time model inference, resulting in substantial carbon footprints per query.

The sheer scale of processing required to generate AI-assisted search strategies raises concerns about sustainability, particularly as demand for AI integration in evidence synthesis continues to grow. Given the increasing use of large language models in biomedical research, the long-term environmental impact of AI-assisted searching warrants careful evaluation and the development of mitigation strategies [3].

Beyond energy consumption, the financial costs of AI-assisted systematic reviewing introduce additional trade-offs. While AI tools claim to reduce screening workloads by 50–75%, their impact on overall research efficiency and cost-effectiveness remains poorly quantified. Many AI-driven platforms operate under subscription-based models, shifting financial burdens from manual labor to licensing fees [13]. For researchers in resource-limited settings, these costs may create new barriers to access, potentially widening disparities in evidence synthesis capabilities.

Furthermore, as AI-assisted screening becomes more prevalent, concerns about data privacy have also emerged. Large language models process large volumes of potentially sensitive information, and their handling of proprietary or confidential research data remains an ethical question. Without robust governance frameworks, the unregulated use of AI for systematic searching may introduce ethical risks, particularly in fields where patient data or unpublished findings are involved [11,12]. The hidden costs of AI in systematic reviewing—environmental, financial, and ethical—highlight the need for a more structured evaluation of its benefits relative to long-term sustainability. Until these trade-offs are better understood, AI should be viewed as an augmentative tool rather than an unequivocal solution for improving research efficiency.

Despite its potential to transform systematic searching, AI remains constrained by critical reproducibility, bias, transparency, and sustainability challenges. The black-box nature of large language models, citation biases, and their environmental impact underscores the importance of cautious and responsible AI integration in evidence synthesis. Ensuring rigorous validation, standardized evaluation frameworks, and adherence to ethical considerations will be crucial in balancing the benefits and limitations of AI-assisted systematic reviews [5].

Until AI can provide reproducible, transparent, and unbiased search strategies, its role in evidence synthesis must remain supplementary. The hybrid AI-human model, where AI enhances efficiency, but human oversight ensures methodological integrity, remains the most viable path forward in the responsible adoption of AI for systematic reviewing [11,12].

6. AI in Evidence Synthesis: A Future that Needs Standards

6.1. AI for living systematic reviews

AI is poised to transform evidence synthesis by enabling living systematic reviews—dynamic, continuously updated reviews that integrate new findings as they emerge [3]. Unlike traditional systematic reviews, which become obsolete as new studies accumulate, AI-driven approaches promise to automate identifying, assessing, and integrating fresh evidence, thereby reducing the need for periodic manual updates [9].

Tools like the EPPI-Reviewer ML classifiers and Covidence's AI-assisted workflows illustrate how machine learning can accelerate the screening and updating process, ensuring systematic reviews remain current and comprehensive in real-time [5]. However, keeping reviews continuously updated introduces unique challenges. AI-assisted updates must preserve methodological rigor, ensuring that newly integrated studies meet predefined quality thresholds and that automated processes do not inadvertently introduce bias [14].

By reducing the time lag between the publication of new research and its inclusion in systematic reviews, AI-driven updates could revolutionize rapidly evolving fields, such as infectious diseases, oncology, and precision medicine. For example, during the COVID-19 pandemic, where hundreds of studies emerged weekly, AI-driven living reviews could have significantly accelerated evidence synthesis for real-time clinical decision-making [11,12]. Similarly, in cancer research, where therapeutic advancements occur at an unprecedented pace, AI-assisted living reviews could help clinicians and policymakers access the latest findings without waiting for traditional review cycles to complete.

However, achieving this vision requires more than automation. AI-generated updates must adhere to the same methodological standards of transparency, reproducibility, and bias mitigation as traditional systematic reviews [13]. Without rigorous validation protocols, automated updates risk introducing inconsistencies or failing to detect methodological flaws in newly included studies.

Additionally, AI models still struggle with key aspects of evidence synthesis, including:

- **Assessing study quality:** AI cannot evaluate study design flaws, inappropriate statistical methods, or confounding factors with the same precision as humans [10].
- **Evaluating the risk of bias:** While some models can flag potential biases, they often misinterpret methodological nuances, requiring expert adjudication [14].
- **Applying nuanced inclusion criteria:** AI systems often overlook subtle distinctions in eligibility criteria, resulting in inconsistent study selection [5].

Until researchers address these limitations, AI should be viewed as an augmentative tool rather than a fully autonomous reviewer. The most effective model remains a hybrid system, where AI accelerates the process, but human experts ensure accuracy and methodological integrity [9].

6.2. Defining an AI validation framework for systematic searching

Despite the growing promise of AI in systematic searching, the field lacks a structured validation framework to ensure AI-driven searches meet the rigorous standards of evidence synthesis. Unlike traditional systematic review methodologies, which emphasize reproducibility, transparency, and relevance, AI-generated searches lack universally accepted benchmarks for accuracy and reliability, raising concerns about consistency, bias, and methodological integrity [11,12].

Researchers must develop a comprehensive AI validation framework to bridge this gap, establishing clear guidelines for assessing AI-generated searches (**Table 4**).

This framework should include:

1. ***Reproducibility safeguards***

○ AI-assisted searches often lack reproducibility, as identical prompts can yield different results due to model updates, variations in prompt structuring, or changes in background training data [10].

○ Traditional Boolean searches offer complete transparency, allowing researchers to precisely track, refine, and validate search queries. In contrast, AI-driven searches remain inherently unstable, leading to inconsistent literature retrieval [14].

○ Manual audits of AI-generated search strategies—where queries are replicated across different time points and evaluated for consistency—must become standard practice to detect query variability and improve reliability [5].

2. ***Sensitivity-specificity trade-off thresholds***

○ AI models often prioritize recall—capturing the broadest set of relevant studies—but this approach introduces high false-positive rates, forcing researchers to screen many irrelevant papers manually [3].

○ Without predefined sensitivity-specificity trade-off thresholds, AI-generated search results may overburden researchers rather than streamline the process [13].

○ Establishing benchmarks for balancing recall and precision—customized for different types of systematic reviews—would help optimize AI-generated searches, ensuring they minimize unnecessary workload while maintaining methodological rigour [11,12].

3. ***Transparency and explainability standards***

○ One of the most significant barriers to AI adoption in systematic reviews is the black-box nature of AI models. Unlike Boolean queries, which provide clear, auditable logic, AI-generated searches rely on non-transparent algorithms that obscure the rationale behind study selection [9].

○ Mandatory transparency reports for AI-assisted searches—detailing search logic, inclusion criteria, and decision-making processes—should become standard, allowing researchers to audit AI outputs and understand why studies were included or excluded [10].

○ Without precise mechanisms for explainability, AI-driven searches risk introducing hidden biases, favoring well-cited studies, specific journals, or high-impact research while overlooking emerging evidence from smaller or less visible sources [14].

4. ***Ethical and environmental considerations***

- AI-assisted systematic searching is a computational and environmental challenge. LLM-powered searches consume four to five times more energy than traditional database queries, raising concerns about sustainability and responsible AI deployment [3].
- As AI models become more prevalent in evidence synthesis, their financial and accessibility costs must also be taken into account. Many AI-powered systematic review platforms operate under subscription-based models, which may potentially widen disparities in access for researchers in low-resource settings [13].
- Future AI validation frameworks should incorporate ethical guidelines, ensuring that AI models do not exacerbate inequalities in scientific research or increase environmental costs disproportionately [11,12].

Despite its transformative potential, AI in evidence synthesis remains constrained by unresolved challenges related to reproducibility, transparency, and sustainability. The lack of standardized validation frameworks, inconsistent AI outputs, and black-box decision-making continue to limit the role of AI in systematic reviews [5].

For AI to transition from an experimental tool to a reliable asset in evidence synthesis, the field must adopt rigorous validation frameworks, including reproducibility safeguards, sensitivity-specificity benchmarks, transparency reports, and ethical oversight [14].

Until these foundational issues are addressed, AI will remain a supplementary rather than a standalone solution. The hybrid AI-human approach, where AI accelerates literature identification and screening, but humans ensure methodological accuracy and integrity, remains the most viable model for the future of AI in systematic reviews [11,12].

Table 4. AI validation framework for systematic searching.

Validation Criterion	Definition	Implementation Strategy
Reproducibility	AI-generated queries should be stable over repeated runs	Require manual audits and sensitivity testing
Bias Control	AI should mitigate citation bias and selection bias	Implement model explainability tools
Sensitivity-Specificity Trade-off	AI searches should balance false positives and false negatives	Establish performance thresholds before study selection
Transparency	Users must understand AI decision-making	Mandate AI-generated transparency reports

7. Discussion

AI is reshaping the landscape of systematic searching and screening, enabling the rapid processing of vast amounts of scientific literature. However, despite its transformative potential, AI remains a tool that extends rather than replaces human expertise. Large language models and machine learning classifiers provide solutions to alleviate the burden of manual screening, identifying relevant studies with sensitivity levels comparable to those of human reviewers [3]. However, AI lacks the nuanced judgment necessary for systematic reviews and cannot function as an autonomous decision-maker. The persistent challenges of reproducibility, bias, and transparency limit its reliability, reinforcing the need for responsible integration rather than blind automation [10].

At the heart of AI-assisted systematic searching lies the convergence of machine learning, deep learning, and natural language processing. Unlike Boolean searches, which operate under explicitly defined rules and produce verifiable outputs, AI-driven models function probabilistically, drawing on semantic similarity and pattern recognition rather than structured syntactic logic [14]. This dynamic approach enables AI to refine search queries in real-time but also introduces unpredictability, as identical prompts may generate different search strategies across multiple attempts. The inability to produce stable, reproducible search results compromises the integrity of AI-assisted systematic reviews, necessitating rigorous validation frameworks to ensure consistency and methodological robustness. Systematic reviews risk becoming irreproducible without a structured approach to verifying AI-generated queries, undermining confidence in evidence synthesis [11,12].

Reproducibility is not the only limitation. AI-driven tools often introduce bias, inheriting the systemic imbalances in the scientific literature from which they draw their information. Citation bias is particularly problematic, as AI models disproportionately prioritize highly cited studies and established journals while overlooking emerging research from smaller institutions and underrepresented disciplines [5]. This self-reinforcing cycle amplifies dominant perspectives in the scientific literature, limiting the diversity of included studies and posing a significant challenge for systematic reviews that aim to provide a comprehensive and balanced synthesis of evidence. AI models prioritize well-cited research and studies from high-impact journals, often at the expense of emerging findings, regional studies, or research from less established institutions [13]. Efforts to mitigate these biases through contextual ranking and adaptive weighting have shown promise; however, AI remains fundamentally constrained by the data on which it is trained. Without targeted interventions—such as adjusting algorithms to account for underrepresented studies or incorporating structured diversity checks—AI-assisted searches risk perpetuating an incomplete and skewed representation of the available evidence.

A lack of transparency further complicates the role of AI in systematic reviews. Traditional Boolean searches allow researchers to define and document their search parameters explicitly, creating an auditable and reproducible logic for study selection. In contrast, AI-driven searches function as opaque black-box models, making probabilistic decisions without clear interpretability [9]. This lack of explainability raises concerns about accountability, particularly in fields where systematic reviews inform clinical guidelines, regulatory approvals, and public policy. If AI-generated research influences high-stakes decisions, researchers must have a transparent understanding of how and why certain studies are included or omitted. Current AI models do not yet provide the level of interpretability required for evidence synthesis, reinforcing the need for human oversight to ensure the integrity of AI-assisted systematic reviews [10].

Beyond methodological concerns, AI's computational and environmental costs present additional challenges. Large-scale language models require substantial computational power, consuming four to five times more energy than traditional database searches [3]. The environmental impact of AI-driven systematic reviews remains poorly understood, particularly as the demand for automation in evidence synthesis grows. The financial barriers posed by AI-assisted search and screening platforms, which operate under subscription-based models, further complicate accessibility, particularly for researchers in low-resource settings [11,12]. While AI has the potential to democratize systematic searching by making complex methodologies more accessible, its sustainability and economic implications must be addressed to prevent the inequitable distribution of AI-driven research capabilities.

Despite these challenges, AI remains a valuable tool when integrated responsibly within a hybrid framework combining automation and human expertise. The future of AI in systematic reviews depends on developing robust validation frameworks that ensure reproducibility, transparency, and methodological integrity [5]. Standardized benchmarking systems must be established to assess AI-generated search strategies against best practices in evidence synthesis. Sensitivity-specificity trade-offs should be defined to balance recall and precision, reducing false positives while maintaining comprehensive search coverage [13]. Transparency reports documenting AI search logic and inclusion-exclusion decisions should become standard practice, providing the interpretability needed to verify and refine AI-generated results.

The next phase of AI integration into systematic reviews must prioritize trust and methodological rigor over speed and automation. AI-driven tools should not displace human reviewers but function as intelligent assistants that enhance efficiency while preserving scientific integrity (**Figure 5**). Moving forward, interdisciplinary collaboration among systematic reviewers, data scientists, and domain experts will be crucial in refining AI methodologies, ensuring that automation enhances rather than compromises the principles of evidence-based medicine. AI has the potential to revolutionize systematic searching and screening; however, a commitment to accuracy, transparency, and reproducibility must guide this transformation. AI should not be viewed as an independent arbiter of scientific evidence but as a trusted ally in the evolving landscape of knowledge synthesis [11,12].

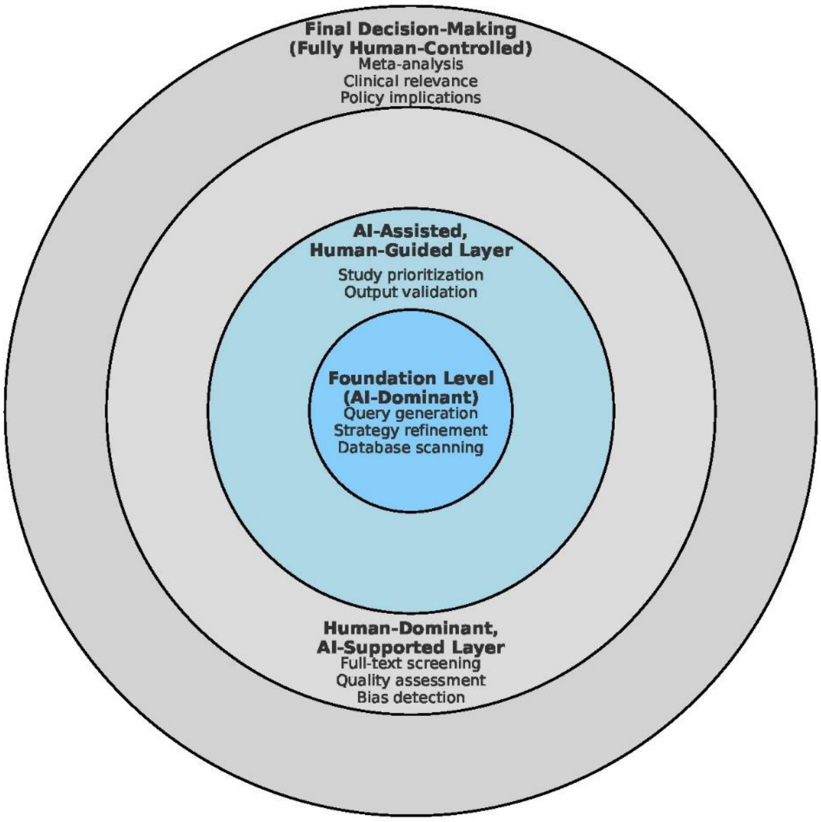

Figure 5. Layered model of AI–human synergy in systematic reviews. A concentric model depicting the progressive interplay between AI systems and human expertise in systematic review workflows. AI dominates the early stages at the foundation, including query construction, search strategy refinement, and database scanning. In mid-level layers, AI assists with study ranking and bias detection, while humans validate outputs and filter irrelevant studies. Human oversight becomes increasingly critical in upper layers, guiding full-text evaluation, assessing methodological rigor, and ensuring clinical relevance. At the apex, the final synthesis and interpretation remain fully human-controlled. This model highlights that while AI can accelerate and enhance review processes, expert judgment remains indispensable for maintaining scientific integrity.

Author Contributions: M.T and O.A.-C.; writing—review and editing. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

Acknowledgments: During the preparation of this manuscript, the authors used Google Collaboration to assist in generating preliminary graphical elements for the figures. The authors subsequently reviewed, edited, and refined these elements using Inkscape (version 1.4.2). The authors have reviewed and edited the output and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
ML	Machine Learning
DL	Deep Learning
LLM	Large Language Model
NLP	Natural Language Processing
MeSH	Medical Subject Headings
SR	Systematic Review
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses
RAG	Retrieval-Augmented Generation

References

1. Adam, G.P.; DeYoung, J.; Paul, A.; Saldanha, I.J.; Balk, E.M.; Trikalinos, T.A.; Wallace, B.C. Literature search sandbox: a large language model that generates search queries for systematic reviews. *JAMIA Open* **2024**, *7*, doi:10.1093/jamiaopen/ooae098.
2. Badami, M.; Benatallah, B.; Baez, M. Adaptive search query generation and refinement in systematic literature review. *Information Systems* **2023**, *117*, doi:10.1016/j.is.2023.102231.
3. Blaizot, A.; Veettil, S.K.; Saidoung, P.; Moreno-Garcia, C.F.; Wiratunga, N.; Aceves-Martins, M.; Lai, N.M.; Chaiyakunapruk, N. Using artificial intelligence methods for systematic review in health sciences: A systematic review. *Res Synth Methods* **2022**, *13*, 353-362, doi:10.1002/jrsm.1553.
4. Cierco Jimenez, R.; Lee, T.; Rosillo, N.; Cordova, R.; Cree, I.A.; Gonzalez, A.; Indave Ruiz, B.I. Machine learning computational tools to assist the performance of systematic reviews: A mapping review. *BMC Med Res Methodol* **2022**, *22*, 322, doi:10.1186/s12874-022-01805-4.
5. Delgado-Chaves, F.M.; Jennings, M.J.; Atalaia, A.; Wolff, J.; Horvath, R.; Mamdouh, Z.M.; Baumbach, J.; Baumbach, L. Transforming literature screening: The emerging role of large language models in systematic reviews. *Proc Natl Acad Sci U S A* **2025**, *122*, e2411962122, doi:10.1073/pnas.2411962122.
6. Feng, Y.; Liang, S.; Zhang, Y.; Chen, S.; Wang, Q.; Huang, T.; Sun, F.; Liu, X.; Zhu, H.; Pan, H. Automated medical literature screening using artificial intelligence: a systematic review and meta-analysis. *J Am Med Inform Assoc* **2022**, *29*, 1425-1432, doi:10.1093/jamia/ocac066.
7. Guo, E.; Gupta, M.; Deng, J.; Park, Y.J.; Paget, M.; Naugler, C. Automated Paper Screening for Clinical Reviews Using Large Language Models: Data Analysis Study. *J Med Internet Res* **2024**, *26*, e48996, doi:10.2196/48996.
8. Lieberum, J.L.; Tows, M.; Metzendorf, M.I.; Heilmeyer, F.; Siemens, W.; Haverkamp, C.; Bohringer, D.; Meerpohl, J.J.; Eisele-Metzger, A. Large language models for conducting systematic reviews: on the rise, but not yet ready for use - a scoping review. *J Clin Epidemiol* **2025**, 10.1016/j.jclinepi.2025.111746, 111746, doi:10.1016/j.jclinepi.2025.111746.
9. Oami, T.; Okada, Y.; Nakada, T.A. Performance of a Large Language Model in Screening Citations. *JAMA Netw Open* **2024**, *7*, e2420496, doi:10.1001/jamanetworkopen.2024.20496.
10. Tran, V.T.; Gartlehner, G.; Yaacoub, S.; Boutron, I.; Schwingshackl, L.; Stadelmaier, J.; Sommer, I.; Alebouyeh, F.; Afach, S.; Meerpohl, J., et al. Sensitivity and Specificity of Using GPT-3.5 Turbo Models for Title and Abstract Screening in Systematic Reviews and Meta-analyses. *Ann Intern Med* **2024**, *177*, 791-799, doi:10.7326/M23-3389.
11. Page, M.J.; McKenzie, J.E.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E., et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ* **2021**, *372*, n71, doi:10.1136/bmj.n71.
12. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E., et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160, doi:10.1136/bmj.n160.

13. Khalil, H.; Ameen, D.; Zarnegar, A. Tools to support the automation of systematic reviews: a scoping review. *J Clin Epidemiol* **2022**, *144*, 22-42, doi:10.1016/j.jclinepi.2021.12.005.
14. Parisi, V.; Sutton, A. The role of ChatGPT in developing systematic literature searches: an evidence summary. *Journal of EAHIL* **2024**, *20*, 30-34, doi:10.32384/jeahil20623.
15. O'Connor, A.M.; Tsafnat, G.; Thomas, J.; Glasziou, P.; Gilbert, S.B.; Hutton, B. A question of trust: can we build an evidence base to gain trust in systematic review automation technologies? *Syst Rev* **2019**, *8*, 143, doi:10.1186/s13643-019-1062-0.
16. Staudinger, M.; Kusa, W.; Piroi, F.; Lipani, A.; Hanbury, A. A Reproducibility and Generalizability Study of Large Language Models for Query Generation. In Proceedings of Proceedings of the 2024 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region; pp. 186-196.
17. Tercero-Hidalgo, J.R.; Khan, K.S.; Bueno-Cavanillas, A.; Fernandez-Lopez, R.; Huete, J.F.; Amezcua-Prieto, C.; Zamora, J.; Fernandez-Luna, J.M. Artificial intelligence in COVID-19 evidence syntheses was underutilized, but impactful: a methodological study. *J Clin Epidemiol* **2022**, *148*, 124-134, doi:10.1016/j.jclinepi.2022.04.027.
18. Wilkins, D. Automated title and abstract screening for scoping reviews using the GPT-4 Large Language Model. *arXiv* **2023**, https://doi.org/10.48550/arXiv.2311.07918, 2311.07918v07911, doi:https://doi.org/10.48550/arXiv.2311.07918.
19. Khraisha, Q.; Put, S.; Kappenberg, J.; Warraitch, A.; Hadfield, K. Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages. *Res Synth Methods* **2024**, *15*, 616-626, doi:10.1002/jrsm.1715.
20. Li, M.; Sun, J.; Tan, X. Evaluating the effectiveness of large language models in abstract screening: a comparative analysis. *Syst Rev* **2024**, *13*, 219, doi:10.1186/s13643-024-02609-x.
21. Abogunrin, S.; Muir, J.M.; Zerbini, C.; Sarri, G. How much can we save by applying artificial intelligence in evidence synthesis? Results from a pragmatic review to quantify workload efficiencies and cost savings. *Front Pharmacol* **2025**, *16*, 1454245, doi:10.3389/fphar.2025.1454245.
22. Matsui, K.; Utsumi, T.; Aoki, Y.; Maruki, T.; Takeshima, M.; Takaesu, Y. Human-Comparable Sensitivity of Large Language Models in Identifying Eligible Studies Through Title and Abstract Screening: 3-Layer Strategy Using GPT-3.5 and GPT-4 for Systematic Reviews. *J Med Internet Res* **2024**, *26*, e52758, doi:10.2196/52758.
23. Fenske, R.F.; Otts, J.A.A. Incorporating Generative AI to Promote Inquiry-Based Learning: Comparing Elicit AI Research Assistant to PubMed and CINAHL Complete. *Med Ref Serv Q* **2024**, *43*, 292-305, doi:10.1080/02763869.2024.2403272.
24. Waffenschmidt, S.; Sieben, W.; Jakubeit, T.; Knelangen, M.; Overesch, I.; Buhn, S.; Pieper, D.; Skoetz, N.; Hausner, E. Increasing the efficiency of study selection for systematic reviews using prioritization tools and a single-screening approach. *Syst Rev* **2023**, *12*, 161, doi:10.1186/s13643-023-02334-x.
25. Lieberum, J.L.; Toews, M.; Metzendorf, M.I.; Heilmeyer, F.; Siemens, W.; Haverkamp, C.; Bohringer, D.; Meerpohl, J.J.; Eisele-Metzger, A. Large language models for conducting systematic reviews: on the rise, but not yet ready for use-a scoping review. *J Clin Epidemiol* **2025**, *181*, 111746, doi:10.1016/j.jclinepi.2025.111746.
26. Adam, G.P.; DeYoung, J.; Paul, A.; Saldanha, I.J.; Balk, E.M.; Trikalinos, T.A.; Wallace, B.C. Literature search sandbox: a large language model that generates search queries for systematic reviews. *JAMIA Open* **2024**, *7*, ooae098, doi:10.1093/jamiaopen/ooae098.
27. Thomas, J.; McDonald, S.; Noel-Storr, A.; Shemilt, I.; Elliott, J.; Mavergames, C.; Marshall, I.J. Machine learning reduced workload with minimal risk of missing studies: development and evaluation of a randomized controlled trial classifier for Cochrane Reviews. *J Clin Epidemiol* **2021**, *133*, 140-151, doi:10.1016/j.jclinepi.2020.11.003.
28. Siemens, W.; von Elm, E.; Binder, H.; Bohringer, D.; Eisele-Metzger, A.; Gartlehner, G.; Hanegraaf, P.; Metzendorf, M.I.; Mosselman, J.J.; Nowak, A., et al. Opportunities, challenges and risks of using artificial

intelligence for evidence synthesis. *BMJ Evid Based Med* **2025**, 10.1136/bmjebm-2024-113320, doi:10.1136/bmjebm-2024-113320.

29. Markolf, S.A.; Chester, M.V.; Allenby, B. Opportunities and Challenges for Artificial Intelligence Applications in Infrastructure Management During the Anthropocene. *Frontiers in Water* **2021**, *2*, doi:10.3389/frwa.2020.551598.
30. Park, S.E.; Thomas, J. Evidence synthesis software. *BMJ Evid Based Med* **2018**, *23*, 140-141, doi:10.1136/bmjebm-2018-110962.
31. Page, M.J.; Moher, D.; Bossuyt, P.M.; Boutron, I.; Hoffmann, T.C.; Mulrow, C.D.; Shamseer, L.; Tetzlaff, J.M.; Akl, E.A.; Brennan, S.E., et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. *BMJ* **2021**, *372*, n160, doi:10.1136/bmj.n160.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.